

Mazsola – mindenkinek

Sass Bálint

MTA Nyelvtudományi Intézet
sass.balint@nytud.mta.hu

Kivonat A igék bővítményszerkezetének vizsgálatára szolgáló *Mazsola* eszközt szabadon elérhetővé tesszük. Ezzel megnyílik a lehetőség, hogy a saját adatai használatával bárki egyszerűen készítsen hasonló funkcionálisú kutatóeszközt. Amint látni fogjuk, a Mazsola alkalmas nem csak magyar nyelvű, hanem más nyelvű szövegek kezelésére is; nem csak egynyelvű, hanem párhuzamos korpuszra is; sőt nem csak igék és bővítmények, hanem más struktúrák vizsgálatára is.

Kulcsszavak: korpusz, ige, bővítmény, többszavas kifejezés, korpuszlekérdező

1 Bevezetés

Az igék és bővítmények vizsgálatára szolgáló *Mazsola* [1,2,3] korpuszlekérdező eszköz felülete régóta hozzáférhető online. A szöveganyag, amin a keresés történik a Magyar Nemzeti Szövegtár első, 187,6 millió szavas változatának [4] egy korábban már közzétett speciálisan a Mazsola számára feldolgozott formája [5].

Az elmúlt évek során a Mazsola magyar szakmai körökben viszonylag ismertté vált. Használják a nyelvtechnológiában [6,7], kontrasztív nyelvészeti [8] és nyelvelméleti kutatásokban [9,10], valamint a magyar mint idegen nyelv oktatása során [11,12]. 2013-ban született egy szakdolgozat, mely lengyel szövegeket feldolgozva a lengyel igéket vizsgálja hasonló felfogásban [13], és keletkeztek olyan igei vonzatkeret- illetve bővítménykeret-adatbázisok, melyek egyik ihletőjének a Mazsola tekinthető [14,15].

Maga a szoftvereszköz azonban – általánossága révén – sokkal többre képes annál, mint hogy magyar igék és bővítmények viszonyait kutassuk vele. Felismerve a nyílt hozzáférés előnyeit most közzétesszük a forráskódot. A cikk két fő üzenete tehát a következő: 1) a Mazsola szoftvereszköz mostantól hozzáférhető, bárki által használható, így saját adataink könnyen vizsgálhatók a Mazsola módszerével; 2) megfelelő adatbázis használatával számos egyéb célra is alkalmazható azon túl, ami az eredeti online változathoz közvetlenül látszik. A továbbiakban a szoftver jellemzőiről valamint a kivonatban vázlatosan már említett felhasználási lehetőségekről lesz szó számos példával illusztrálva.

2 Használata, működése

A <http://corpus.nytud.hu/mazsola> címen elérhető magyar nyelvű szövegeket tartalmazó eredeti Mazsola felület alapvetően arra szolgál, hogy feltérképezzük az

egyes magyar igék melletti adott esetraggal/névutóval járó jellegzetes névszók, illetve fordítva: az egyes (esetragos/névutós) névszók mellett jellegzetesen előforduló igéket. (Alapvető használatáról [3] 3.2. fejezetében olvashatunk. Megjegyezzük, hogy a mögöttes adatbázis az elváló igekötőket az igékhez illesztve tartalmazza, így az eszköz lehetőséget biztosít egy ige összes előfordulásának egységes vizsgálatára függetlenül attól, hogy az igekötő éppen elválízik vagy nem.)

Az első esetben – az igék bővítményeire kérdezve – általában azt az eredményt kapjuk, hogy adott ige adott bővítményeként előforduló jellegzetes névszók egy szemantikailag koherens szóosztályt alkotnak, amiben kakukktójásként jelennek meg az idiomatikus, az igével intézményesült kifejezést, komplex igét, szólást alkotó szavak, mint például az *emelkedik -rA*-ragos bővítményei közül kilógó *jogerő*. A skála egyik szélén azok az esetek vannak, mikor szinte egységes a kijövő szóosztály (pl. a *fogyaszt* tárgyai különféle erőforrások); a másik szélén pedig azok, mikor az ige szinte minden egyes jellegzetes névszóval különálló komplex igét alkot (pl. a *hagy* legjellemzőbb tárgyaiból adódó kifejezések: (*sok*) *kivánnivalót hagy maga után*, (*mély*) *nyomot hagy (vhol)*, (*nem*) *hagy kétséget vmi felől*, *hagyja magát*, *üzenetet hagy*). A fentiek különféle variációi is előfordulnak, a *kitör* alanyai például nagyjából két osztályra bomlanak: a konkrét (*vihar*, *vulkán*) és a metaforikus jelentés mentén (*pánik*, *háború*, *botrány*, *forradalom*). Speciális eset, mikor adott ige mellett adott esetraggal csak néhány, szélső esetben egyetlen szó szerepel (pl.: a *csóvál* tárgyaként a *fej*, *farok*, *kutya* vagy a *hunyt* tárgyaként a *szem*).

Fontos kiemelni a második esetet (a fenti eset duálisát), amikor bizonyos bővítmények megadásával az igeire kérdezzük rá. Erről kevesebb szó esett korábban. Általában egyszerűen mindenféle jellegzetesen az adott bővítménnyel használatos igét kapunk eredményül egy ilyen lekérdezésre. A bővítményt megadhatjuk esetrag nélkül is, így esetragtól függetlenül, értsd tetszőleges esetraggal rákérdez a rendszer. A *lő* bővítményre legjellegzetesebbként kapjuk: *ül lovon*, *megül lovat*, *húz ló*, *leszáll lóról*. Bizonyos esetekben itt is szemantikailag koherens szóosztályt kapunk: nagyon hasonló jelentésű igéket. Erre talán legszebb példa a *hideg + hátán* bővítményekre kapott *végigfut*, *futkos*, *futkározik* igeosztály (1. ábra). A *kígyót* bővítmény két legjellegzetesebb igéje viszont a *kiabál* és a *melenget*, melyekkel ez a szó külön-külön önálló szerkezetet, szólást alkot. És itt is előfordul az az eset, hogy kizárólag egy ige szeret az adott bővítménnyel társulni (pl.: *vérszemmet* → *kap*, vagy *füst alatt* → *elintéz*).

A Mazsolánál jóval összetettebb, de mégis hozzá sok szempontból hasonló eszköz a közismert SketchEngine [16,17]. A SketchEngine-ben – a benne alkalmazott szkeccs-nyelvtanok (sketch grammar) segítségével – explicite megmondjuk, hogy mely elemek kapcsolataira vagyunk kíváncsiak, a Mazsolában ezzel szemben implicite minden mindenhez kapcsolódik, és az összes ilyen kapcsolat vizsgálható. A SketchEngine alapvetően mindig két elem viszonyát nézi. Van benne kiegészítő eszköz, amivel meg lehet valósítani 3 elem összekapcsolását, de ez az, amiben a Mazsola jobb: *rekurzívan* képes a korábban összekapcsolt elemekhez tetszőleges számú újat kapcsolni, és egy kiválasztott elemnek a jellegzetességét vizsgálni a kialakult szerkezet többi tagja összességéhez képest, amint ezt a fenti példák-



Korpusz: Magyar Nemzeti Szövegtár

Igétő:

Nem: <input type="checkbox"/> Eset/névutó: alany	Nem: <input type="checkbox"/> Vonzattó: hideg
Nem: <input type="checkbox"/> Eset/névutó: -n	Nem: <input type="checkbox"/> Vonzattó: hát
Nem: <input type="checkbox"/> Eset/névutó: <input type="text"/>	Nem: <input type="checkbox"/> Vonzattó: <input type="text"/>
Nem: <input type="checkbox"/> Szó: <input type="text"/>	

Teljes mondatlefedés: ☐

Mehet

113 találat. [végigfut](#) [35] [futkos](#) [30] [futkározik](#) [13]

1. ábra: A *hideg* és *hátán* bővítmények megadásának módja, és a hozzájuk tartozó jellegzetes igék lekérdezése. Fontos, hogy az „Eloszlás” kapcsoló az ige sorára van állítva.

ban láttuk is. Megvizsgálhatjuk tehát, hogy a *tesz*-nek milyen jellegzetes tárgyai vannak (*elég*, *szer*, *feljelentés*, *ígéret* stb.), majd hogy az *tesz eleget*-nek milyen *-nAk* ragos bővítményei (*kötelezettség*, *követelmény*, *kérés*, *meghívás* stb.), aztán hogy a *tesz eleget kötelezettségnek* szerkezetnek milyen alanyai (*önkormányzat*, *kormány*, *országgyűlés*, *főváros*), és így tovább. Fontos különbség még, hogy a Mazsola immár szabadon és ingyenesen hozzáférhető.

3 A szoftver jellemzői, elérhetősége, installálása

A Mazsola letölthető a <https://github.com/sassbalint/mazsola> repozitóriumból. Perl nyelven van megvalósítva, linux operációs rendszeren működik.

Az installálás leírását a github-on található **README.md**-ben olvashatjuk. Alapvetően annyit kell tenni, hogy **make** segítségével külön-külön installálni kell magát a szoftvert, illetve az adatbázisokat egy webszerverre. Az installálás után a Mazsola böngészőből használható. A repozitórium tartalmazza a fő szkriptet (**mazsola_skel.pl**), a többnyelvűsítést megvalósító szkriptet, különböző config fájlokat, valamint az installáláshoz szükséges **Makefile**-t.

Ezen kívül itt találjuk még a **data** könyvtárt, ebben helyezhetők el a korpuszok, speciális, a Mazsola számára alkalmas formátumban. A Mazsola-adatbázisok formátumának leírása [5] 2. fejezetében található.

A Mazsola repozitóriuma eleve tartalmaz több kisméretű - néhány ezer tagmondatból álló - minta-adatbázist (1. táblázat). Egyrészt, hogy ezek segítségével ki lehessen próbálni a rendszer működését, másrészt, hogy alkalmas példákon tanulmányozhassuk az ilyen adatbázisok szerkezetét, kinézetét.

Az egyes minta-korpuszokhoz az installálás után a Mazsola felületén megfelelő példákat is találunk, melyekre rákattintva és a kérést lefuttatva kipróbálhatjuk az eszközt. A dán korpusz hasonló részletességgel van elemezve, mint a magyar. A szerb korpuszban egy primitívebb megoldás van megvalósítva. Mivel a korpuszban nem volt triviális meghatározni, hogy melyik szó alany és melyik tárgy, az előljáró nélküli főneveket egy kalap alá vettük, és ezek mind **ANYCASE**-ként vannak kódolva. A példák kipróbálásakor láthatjuk, hogy ebből az egyszerűbb kódolással ellátott korpuszból is a korábbiakhoz hasonló hasznos eredményeket

1. táblázat. A Mazsolához adott minta-adatbázisok forrásai.

körpusz	nyelv
MNSZ [4]	magyar
Danish Dependency Treebank [18]	dán
Intera körpusz [19]	szerb
Dutch Parallel Corpus [20]	holland-francia

tud előhozni a Mazsola. Az *imati* (birtokol) ige melletti tárgyakat (az ANYCASE segítségével) nézve jellegzetesként megkapjuk a *pravo* (jog) szót, majd a *na* előljáróval kiegészítve a keresést, hogy mire lehet jogunk.

Látjuk azt, hogy nyugodtan alkalmazhatjuk más nyelvekre az eszközt, ugyanolyan jellegű működést fogunk tapasztalni, mint amit a magyar esetén leírtunk. Természetesen a Mazsola mindig azon információk alapján tud dolgozni, melyek az adatbázisban kódolva vannak, másképp fogalmazva az, hogy adott Mazsola-példány milyen eredményeket ad, attól függ, hogy mit kódoltunk bele az adatbázisba. Alkalmasan választott kódolással szemantikai összefüggéseket vagy párhuzamos körpuszt is kezelni tudunk, utóbbira még visszatérünk.

Ha saját adatbázist akarunk használni, akkor el kell helyezni a **data** könyvtárban, és készíteni kell hozzá egy megfelelő config fájlt (amit – ahogyan ezt a repozitóriumban most is látjuk – `mazsola_config_skel.pl` néven kell belinkelni). A config fájlban be kell állítanunk a körpuszt: meg kell adnunk 1) az adatbázisfájl nevét; 2) a szótőgyakoriságokat tartalmazó kiegészítő adatbázisfájl nevét (amit a `lemmafreq` könyvtárban kell elhelyezni); 3) a körpusz nevét; 4) és azt a küszöbértéket, aminél gyakrabban előforduló legjellegzetesebb szavakat az eredményoldalak felső részén meg szeretnénk jeleníteni. Ha rendelkezésre áll a szótőgyakoriságokat tartalmazó kiegészítő adatbázis, akkor tud a rendszer *jellegzetességet* (*salience*-et) számolni, egyébként csak az egyszerű gyakoriságokat jeleníti meg. A fentiekén kívül meg kell adnunk a körpusz karakterkódolását (jelenleg csak egybájtos karakterkódolások támogatottak), valamint megfelelő formátumban a felületen megjelenítendő, körpuszhoz tartozó kattintható példákat. Végül el kell végezni az installálás lépéseit. Config fájlra is találunk néhány mintát a repozitóriumban (`mazsola_config_*.pl`).

4 Létező Mazsolák

A minta-adatbázisoknak megfelelő teljes körpuszok nagy része elérhető a Mazsolán keresztül online a 2. táblázatban szereplő internetes címeken.

Ezek a felületeken is találunk példákat, illetve a minta-körpuszokhoz tartozó példák kipróbálhatók ezeken az URL-eken is jóval nagyobb méretű szövegyanyagon.

2. táblázat. A Mazsolához adott minta-adatbázisok forrásai.

körpusz	internetes cím
MNSZ	http://corpus.nytud.hu/mazsola
Danish Dependency Treebank	http://corpus.nytud.hu/vabd [21]
Intera körpusz	http://corpus.nytud.hu/vabs
Dutch Parallel Corpus	http://corpus.nytud.hu/vabnlfr (terv) [22]

5 Összetettebb példák, trükkök, kiterjesztések

5.1 Egy korábbi példa

Korábban, egy az elemzetlen Magyar történelmi szövegtáron végzett egyszerű vizsgálat kapcsán azt írtam, hogy az *erőt vesz rajta* alanyai általában „érzések, azon belül is inkább negatív tartalmú érzések, hangulatok” [23]. Nézzük most ezt meg a Mazsolán, ami kifejezetten az effajta kérdések megválaszolására szolgál. Azt kapjuk, hogy az *erőt vesz* kifejezés *-n* ragos jellegzetes bővítményei: *maga*, *ő*, *ők*, *mi*, *ember*. A *maga*-t – más kifejezés lévén – érdemes a mostani vizsgálatból kihagyni, a többi relevánsnak tűnik: megadjuk tehát esetnek a *-n* ragot, és vonzattövek közül letiltjuk („Nem” jelölőnégyzet) a *maga*-t. Ennek a kifejezésnek (*erőt vesz vmin* (kivéve: *magán*)) kérdezzük le a jellegzetes alanyait a Mazsola fentebb leírt rekurzív vizsgálati módszerét alkalmazva (2. ábra).

Mazsola - a magyar igei bővítményszerkezet vizsgálata

English version

Körpusz: Magyar Nemzeti Szövegtár

Igető: vesz

Nem: ☐ Eset/névutó: t Nem: ☐ Vonzattő: erő

Nem: ☐ Eset/névutó: n Nem: ☒ Vonzattő: maga

Nem: ☐ Eset/névutó: alany Nem: ☐ Vonzattő:

Nem: ☐ Szó:

Teljes mondatlefedés: ☐

Mehet

198 találat. [érzés](#) [8] [kapu](#) [7]

Eloszlás:

- ☐
- ☐
- ☐
- ☒

2. ábra: Az *erőt vesz vmin* (kivéve: *magán*) szerkezet jellegzetes alanyai. „Eloszlás” kapcsoló az alany során.

Azt kapjuk, hogy a legjellegzetesebb alany éppen az *érzés*, és a további találatok között megtaláljuk a *bágyadtság*, *bénaság*, *betegség*, *borzongás*, *búskomorság*, *bűntudat* stb. szavakat, azaz az eredeti állítás megállja a helyét. A második legjellegzetesebb alany viszont a *kapu*, itt egy (bibliai) szólás bukkan fel kakukktójásként, a konkrét vizsgálatból a *maga*-hoz hasonlóan ezt is érdemes kihagyni.

5.2 Reguláris kifejezések

Megjegyezzük, hogy a lekérdezőfelület mezőiben bizonyos reguláris kifejezéseket is használhatunk.

Ha például az a feladat, hogy *ezik/szik* alternációt keressünk az *emlékezik/emlékszik* példájára, akkor az alapján a sejtés alapján, hogy az ilyen igéknek *ék*-végű főnévből kell származnia, megpróbálkozhatunk az

Igető=. *ékezik

lekérdezéssel. Az eredmény: *emlékezik, verejtékezik*. Ezek valóban *ék*-végű főnévből származnak, de *verejtékszik* alak ennek ellenére nincsen. Úgy tűnik, hogy az *emlékezik* ebből a szempontból egyedi. Hasonlóan az

Igető=. *g[ae]t

lekérdezés segítségével megragadhatjuk a *-gAt*-képzős igéket.

A „Vonzattő” és a „Szó” mező esetén több szóalakot is megadhatunk, ez úgy értelmeződik, hogy a szólista bármelyik tagja előfordulhat. Azaz ha például férfiakra jellemző igéket keresünk, akkor megadhatjuk a alanyként a *férfi, fiú, kamasz, kisfiú, öregember, férj, após, úr* stb. szólistát. Annak ellenére, hogy a szólisták a fentiek szerint diszjunktívan értelmeződnek, meg tudunk adni a „Szó” mezőben többszavas kifejezést is. Reguláris kifejezés cseles használatával, ha a szavakat ponttal kötjük össze, az azt fogja eredményezni, hogy közvetlenül egymás után kell következniük. A *rázza.az.ág* lekérdezés esetén csak azokat a találatokat kapjuk meg, amiben a *rázza az ág* sztring szerepel, nem pedig *rázza, az, ág* szavak bármelyike.

5.3 Feladatmegoldás Mazsolával

Nézzük néhány konkrét problémát, feladattípust, melyeket a Mazsola segítségével oldhatunk meg.

A *macska* és a *cica* közeli szinonimáknak tűnnek. A Mazsola szerint az előbbi *dorombol, nyávog, ül, megesz, leugrik, eltűnik, nyújtózik*, az utóbbi viszont *elvisz*.

Mit jelent az, hogy *kooptál*? Az ige tárgyainak vizsgálatából levonhatjuk azt a következtetést, hogy a jelentés nagyjából az, hogy bizottságba új tagot vesz fel.

A fordítói munka során arra lehet hasznos a Mazsola, hogy bemutatja a bevett kollokációkat. Ha a *részletes magyarázatot kíván* megfogalmazás nem tűnik ideálisnak, akkor élhetünk a Mazsola javaslatával, amely ez esetben: *részletes magyarázatot igényel*.

A *kiküszöböl* a *csorbát* kifejezés kicsit furcsának hat. Biztos, hogy ezt az igét szoktuk itt használni? Milyen igéket ad a Mazsola a *csorbát* tárgyhoz? Az eredmény a *szenved* és az *ejt* mellett több *kiküszörül* és jóval kevesebb *kiküszöböl*. Persze, a *csorbát* inkább *kiküszörül*jük. Mit történhetett? Az igék felől vizsgálódva azt látjuk, hogy a *kiküszörül* szinte kizárólag a *csorbát* tárggyal fordul elő, a két nagyságrenddel gyakoribb *kiküszöböl* viszont számos különféle tárggyal,

melyeknek gyakori jelentéseleme a valamiféle hiányosság (*hiba, hiányosság, probléma, anomália* stb.). A gyakoribb és nagyon hasonló hangzású kifejezés magába olvasztotta a ritkább, hasonló jelentésű kifejezést. Az állítás alátámasztására persze még érdemes történeti korpuszban utánajárni, hogy melyik kifejezés mikortól adatolható. Hasonlóan vizsgálhatók a *számításon kívül hagy* körüli fejlemények.

Kapcsolatot fedezhetünk fel az *össze* igekötő és a *-vAl* ragos vonzatok között. Az *össze* igekötős igék jelentős része *-vAl* ragos vonzattal jár, másik nagy része viszont e nélkül. Milyen eltéréseket találunk a két csoport között? Az első csoport (jellemzően *-vAl* ragos bővítménnyel): *összefügg, összeköt, összehasonlít, összekever, összevet* stb. A második csoport (jellemzően *-vAl* ragos bővítmény nélkül): *összegyűlik, összetörik, összeállít, összehív, összeszed* stb. Azt látjuk, hogy az első esetben két dolgot állítunk relációba, a második esetben pedig sokat. Az *össze* jelentése az első esetben valamiféle „szembe”, a második esetben pedig valamiféle „egybe”. E felismerésnek által két csoportra osztva strukturálhatjuk az *össze* igekötő jelentéseit (az alapigétől függetlenül!), és mondhatjuk, hogy a *-vAl* az első esetben az igekötőnek a vonzata.

5.4 Párhuzamos korpusz kódolása

Térjünk vissza arra a kérdésre, hogy hogyan kezelhető a Mazsolában párhuzamos korpusz. Az adatbázis kódolásában rejlik a megoldás, a szoftver maga változatlan módon használható. Hogyan valósul meg a már említett holland-francia párhuzamos korpusz esetében a megfelelő Mazsola-adatbázis kódolása? Az egy nyelvű magyar `hnc_sample.mazsoladb` mintafájlban ezt látjuk:

```
Ha a város a tárcáktól pénzt nyer , stem@@nyer ABL@@tárca
ACC@@pénz NOM@@város
```

A sor elején maga a tagmondat látható, `stem@@` után következik az igető, majd sorra az esetek a megfelelő szóval együtt. A holland-francia párhuzamos `data/tltnlfr_sample.mazsoladb` mintafájlban viszont a következő sort találjuk:

```
nl_deze te vertalen in wiskundige formules . / fr_les traduire
en expressions mathématiques . stem@@nl_vertalen+fr_traduire
fr_en@@expression nl_in@@formule
```

Amint látjuk, azt a megoldást választottuk, hogy a perrel elválasztott, `nl_` és `fr_` címkével azonosított kétnyelvű tagmondat után az ige helyén a holland-francia igepár van (szintén nyelvi címkével ellátva és + jellel összekapcsolva), a bővítmények között pedig a hollandok és a franciák „vegyesen” egy halmazban szerepelnek. Érdemes élesben kipróbálni a TLT minta-korpuszhoz mellékelt példákat.

5.5 Nem csak igére

Vegyük észre, hogy az eszköz tetszőleges olyan struktúrájú objektumok kutatására alkalmas, ahol egy egységen belül van egy központi elem (esetünkben az ige),

és mellette vannak hozzá tartozó elemek (esetünkben a bővítményi névszók), amiket valamiféle kapcsolat (esetünkben az esetragok/névutók) fűz a központi elemhez.

Ha ilyen struktúrájúak az adataink, akkor tudunk hozzá Mazsolát készíteni, bármilyen konkrét tartalmúak is legyenek az adatok.

Vegyük például a jelző+főnév párokat. Ebben az esetben a főnév lesz a központi elem, a jelző a hozzá tartozó elem, és a kapcsolat csupán annak rögzítése lesz, hogy a jelző a főnévhez tartozik. (Az eredeti ige-központú esetben mindig annyiféle kapcsolat volt, ahány esetrag ill. névutó van, itt viszont mindössze egyfajta kapcsolat van.) Ilyen felépítésű adatbázis használatával a Mazsola működésének megfelelően meg fogjuk kapni a főnevek jellegzetes jelzőit, illetve a jelzőkhöz tartozó jellegzetes főneveket.

6 Konklúzió

Szabadon elérhető, használható a Mazsola eszköz. A saját adatait, korpuszát megfelelően előkészítve, és a rendszerbe beleillesztve bárki készíthet igék és bővítmények vizsgálatára alkalmas kutatóeszközt. A [5]-ben közzétett adatbázis használatával előállítható az eredeti magyar Mazsola felület. Az eredeti rendszer reprodukálásánál azonban sokkal izgalmasabb, ha a saját adatainkra építve a fenti ötletek és kiterjesztések alapján készítünk különféle más nyelvekre és/vagy más struktúrákra vonatkozó korpuszlekérdezőket a Mazsola segítségével.

Köszönetnyilvánítás

A munkálatokat az MTA Bolyai János Kutatási Ösztöndíja támogatta (ügyszám: BO/00064/17/1; időtartam: 2017-2020).

Hivatkozások

1. Sass, B.: „Mazsola” – eszköz a magyar igék bővítményszerkezetének vizsgálatára. In: Váradai Tamás (szerk.): Válogatás az I. Alkalmazott Nyelvészeti Doktorandusz Konferencia előadásaiból, MTA Nyelvtudományi Intézet, Budapest (2009) 117–129
2. Sass, B.: Korpusznyelvészeti eszköz a magyar igék bővítményszerkezetének vizsgálatára. In Sinkovics, B., ed.: LingDok 8. – Nyelvész-doktoranduszok dolgozatai. JATEPress, Szeged (2009) 143–155
3. Sass, B.: Igei szerkezetek gyakorisági szótára - egy automatikus lexikai kinyerő eljárás és alkalmazása. PhD thesis, PPKE ITK (2011)
4. Váradai, T.: The Hungarian National Corpus. In: Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC2002), Las Palmas, Spain (2002) 385–389
5. Sass, B.: 28 millió szintaktikailag elemzett mondat és 500000 igei szerkezet. In: XI. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2015), Szeged: JATEPress (2015) 303–308
6. Prószték, G.: A nyelvtechnológia és a magyar nyelvtudomány. Magyar Nyelv **108**(1) (2012) 1–18

7. Indig, B., Laki, L., Gábor, P.: Mozaik nyelvmodell az anagramma elemzőhöz. In: XII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2016), SZTE, Szeged (2016) 260–270
8. Forintos, É.: New Zealand-Hungarian language contact phenomena. In Bátyi, S., Vigh-Szabó, M., eds.: A nyelv: rendszer, használat, alkalmazás. Budapest: Tinta Könyvkiadó (2015) 199–211
9. Kálmán, L.: Bővítménykeretek mint konstrukciók. In Kas, B., ed.: Szavad ne feledd! Tanulmányok Bánréti Zoltán tiszteletére. Budapest: MTA Nyelvtudományi Intézet (2016) 61–72
10. Kenesei, I.: A szófaj fogalmáról és meghatározásáról. In Kas, B., ed.: Szavad ne feledd! Tanulmányok Bánréti Zoltán tiszteletére. Budapest: MTA Nyelvtudományi Intézet (2016) 81–92
11. Joachim, L.: Morfológiai és szintaktikai információk a magyar mint idegen nyelv tanulói számára készített szótárakban és szöszedetekben (egy tanulói szótár előmunkálatainak szempontjai, tapasztalatai). In Szűcs, T., Nádor, O., eds.: Hungarológiai Évkönyv 14. Pécs: Dialóg Campus Kiadó (2013) 40–62
12. Wéber, K.: Elvárt és használt szintaktikai szerkezetek megoszlása b1, b2 és c1 szintű magyar mint idegen nyelvi fogalmazásfeladatokban. In Szűcs, T., Nádor, O., eds.: Hungarológiai Évkönyv 14. Pécs: Dialóg Campus Kiadó (2013) 63–73
13. Wolosz, M.: A lengyel igei szerkezetek számítógépes kinyerése és gyakorisági vizsgálatuk. Master's thesis, ELTE IK (2013)
14. Kornai, A., Nemeskey, D.M., Recski, G.: Detecting optional arguments of verbs. In Chair), N.C.C., Choukri, K., Declerck, T., Goggi, S., Grobelnik, M., Maegaard, B., Mariani, J., Mazo, H., Moreno, A., Odijk, J., Piperidis, S., eds.: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Paris, France, European Language Resources Association (ELRA) (2016)
15. Vadász, N., Kalivoda, A., Indig, B.: Egy egységesített magyar igei vonzatkerettár építése és felhasználása. In: XIV. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2018), SZTE, Szeged (2018) jelen kötetben
16. Kilgariff, A., Rychly, P., Smrz, P., Tugwell, D.: The Sketch Engine. In: Proceedings of EURALEX 2004, Lorient, France (2004) 105–116
17. Kilgariff, A.: The Sketch Engine: Ten years on. *Lexicography* **1**(1) (2014) 7–36
18. Trautner Kromann, M.: The Danish Dependency Treebank and the DTAG treebank tool. In: Proceedings of the 2nd Workshop on Treebanks and Linguistic Theories (TLT 2003), Växjö, Sweden (2003)
19. Gavriliadou, M., Labropoulou, P., Piperidis, S., Giouli, V., Calzolari, N., Monachini, M., Soria, C., Choukri, K.: Language resources production models: the case of the INTERA multilingual corpus and terminology. In: Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC2006). (2006)
20. Macken, L., Trushkina, J., Paulussen, H., Rura, L., Desmet, P., Vandeweghe, W.: Dutch Parallel Corpus. A multilingual annotated corpus. In: Proceedings of Corpus Linguistics 2007, Birmingham, United Kingdom (2007)
21. Sass, B.: Verb Argument Browser for Danish. In: Proceedings of the 17th Nordic Conference of Computational Linguistics, NoDaLiDa 2009, Odense, Denmark (2009) 263–266
22. Sass, B.: Párhuzamos igei szerkezetek közvetlen kinyerése párhuzamos korpuszból. In: VII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2010), SZTE, Szeged (2010) 102–110
23. Sass, B.: Keresés korpuszban: a kibővített Magyar történeti szövegtár új keresőfelülete. In Forgács, T., Németh, M., Sinkovics, B., eds.: A nyelvtörténeti kutatások újabb eredményei IX., Szeged: SZTE Magyar Nyelvészeti Tanszék (2017) 267–277